PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 9/20, G06T 15/70, H04N 7/26 — A1

(11) International Publication Number: WO 97/36288

(43) International Publication Date: 2 October 1997 (02.10.97)

(21) International Application Number: PCT/GB97/00818

(22) International Filing Date: 24 March 1997 (24.03.97)

(30) Priority Data:
96302060.7 26 March 1996 (26.03.96) EP
(34) Countries for which the regional or international application was filed: GB et al.

(71) Applicant *(for all designated States except US)*: BRITISH TELECOMMUNICATIONS PLC [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

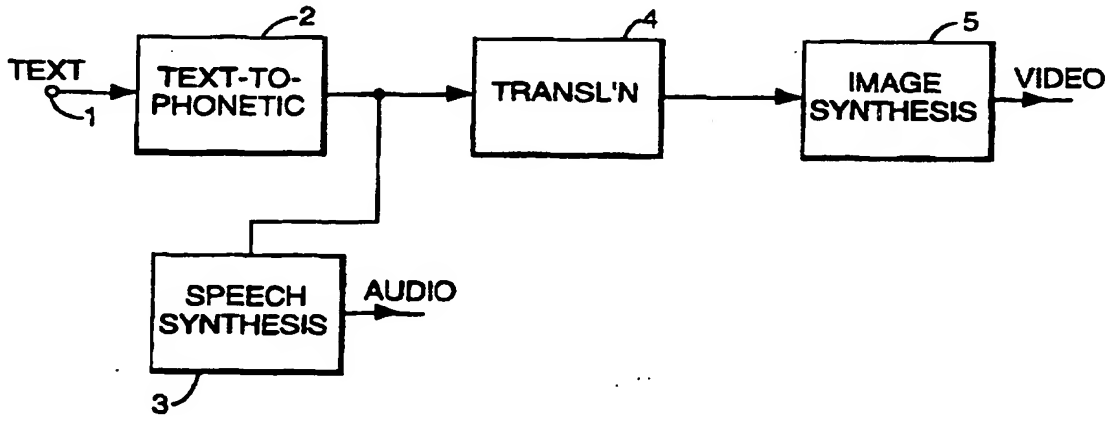
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: BREEN, Andrew, Paul [GB/GB]; 50 Westerfield Road, Ipswich, Suffolk IP4 2UT (GB). BOWERS, Emma, Jane [GB/GB]; 1 Vine Close, Oundle, Peterborough PE8 4A2 (GB).

(74) Agent: LLOYD, Barry, George, William; BT Group Legal Services, Intellectual Property Dept., 8th floor, 120 Holborn, London EC1N 2TE (GB).

(81) Designated States: AU, CA, CN, JP, KR, MX, NO, NZ, SG, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: IMAGE SYNTHESIS



(57) Abstract

Synthesis of a moving picture of a face (e.g. to accompany synthetic speech) is performed by converting an input phoneme string into a sequence of mouth shapes or visemes. Specifically a shape is generated for each vowel and for each transition involving a consonant.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

IMAGE SYNTHESIS

The present invention relates to synthesis of moving images, for example to accompany synthetic speech.

According to the present invention there is provided a method of generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, comprising:

receiving a sequence of phonetic representations corresponding to successive portions of the utterance;

identifying a mouth shape for each phonetic representation of a first type;

identifying a mouth shape for each transition from a phonetic representation of the first type to a phonetic representation of a second type, for each transition from a phonetic representation of the second type to a phonetic representation of a first type and for each transition from a phonetic representation of the second type to a phonetic representation of the second type; and

generating a sequence of image frames including the identified shapes.

The first and second types may be vowels and consonants respectively; thus, a preferred embodiment of the invention provides a method of generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, comprising:

receiving a sequence of phonetic representations corresponding to successive phonemes of the utterance;

identifying a mouth shape for each vowel phoneme;

identifying a mouth shape for each transition from a vowel phoneme to a consonant phoneme, for each transition from a consonant phoneme to a vowel phoneme and for each transition from a consonant phoneme to a consonant phoneme; and

generating a sequence of image frames including the identified shapes.

The identification of a mouth shape for each transition between consonant and vowel phonemes may be performed as a function of the vowel phoneme and the consonant phoneme, whilst the identification of a mouth shape for each transition between two consonant phonemes may be performed as a function of

the first of the two consonant phonemes and of the vowel phoneme which most closely follows or precedes it. Alternatively the identification of a mouth shape for each transition between two consonant phonemes may be performed as a function of the first of the two consonant phonemes and of the vowel phoneme which most closely follows it or in the absence thereof that which precedes it.

Preferably the identification for each transition is performed as a function of only those phonemes specified above in relation to those transitions. Alternatively, the identification could be performed as a function also of at least one other phoneme within the same word.

In a preferred arrangement, one may generate for each identified mouth shape a command specifying that shape and intermediate commands each of which specifies a shape intermediate the shapes specified by the preceding and following commands.

In another aspect of the invention there is provided an apparatus for generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, comprising:

means arranged in operation to receive a sequence of phonetic representations corresponding to successive portions of the utterance and in response thereto to

identify a mouth shape for each phonetic representation of a first type and

identify a mouth shape for each transition from a phonetic representation of the first type to a phonetic representation of a second type, for each transition from a phonetic representation of the second type to a phonetic representation of a first type and for each transition from a phonetic representation of the second type to a phonetic representation of the second type;

and means for generating a sequence of image frames including the identified shapes.

One embodiment of the invention will now be described, by way of example, with reference to the accompanying drawings, in which:

Figure 1 is a functional block diagram showing the elements of the embodiment;

Figure 2 shows a plan, and front and side elevations of the 'wireframe' used in synthesising an image of a human head;

Figure 3 shows similar views of a 'wireframe' used in synthesising an the mouth portion of an image of a human head;

Figure 4 shows where the maximum vowel mouth shapes occur in the synthesis of a sequence of images to represent a human head saying 'affluence';

Figure 5 shows where the maximum vowel to consonant (and vice versa) transitional mouth shapes occur in the word 'affluence';

Figure 6 illustrates the remaining mouth shapes in the articulation of the word 'affluence';

Figure 7 illustrates the transitions between the mouth shapes in the articulation of the word 'affluence';

Figure 8 is a block diagram schematically illustrating the components of the unit for translating phonetic signals into command signals for the image synthesis unit;

Figure 9 is a flow chart illustrating the operation of the apparatus of the embodiment;

Figure 10 is a flow chart illustrating the procedure for conversion of diphthongs and affricates into their constituent phonemes;

Figures 11A to 11D illustrate the procedure for producing an intermediate output file on the basis of the input phoneme file;

Figure 12 illustrates the procedure for producing a file specifying the timing and nature of the maximal mouth shapes on the basis of the intermediate output file; and

Figures 13A and 13B illustrate the procedure for producing a file specifying both the maximal mouth shapes and intermediate mouth shapes.

The apparatus of Figure 1 has the function of receiving words to be spoken, in the form of text, and generating corresponding speech in the form of an audio signal and generating a corresponding video signal for display of a moving picture of a face (human or cartoon for example) with mouth articulation which corresponds to that same speech. In this description, reference will often be made to mouth articulation; it is to be understood that this articulation may include movement of the lips, interior of the mouth (including, if wished, teeth and tongue), jaw, and surrounding areas. Other movements such as gross head movement or rotation, eyebrow movement and so forth may be included also in order to make the resulting image appear more realistic.

The text, from a stored text file or other desired source is received at an input 1 in the form of character codes according to any convenient standard representation (e.g. ASCII code). It is received by a speech synthesiser of conventional construction but shown here as two separate parts, namely a text-to-phonetic converter 2 which converts conventional orthography into a phonetic representation, for example, a list of phonemes and the duration of each, and the speech synthesiser proper 3 which converts the list into an audio frequency waveform. Any phoneme set may be used, but for the purposes of this description use of the British RP-SAMPA set is assumed which identifies 38 distinct phonemes of British English as set out in Table 1 below.

TABLE 1

BRITISH RP-SAMPA	WORD EXAMPLE
CONSONANTS	
/b/	bear
/D/	this
/d/	dear
/f/	fear
/g/	gear
/h/	hear
/j/	year
/k/	king
/l/	lead
/m/	men
/N/	wing
/n/	near
/p/	pear
/r/	rear
/S	sheer
/s/	sing
/T/	thing
/t/	tear
/v/	very
/w/	wear
/Z/	treasure
/z/	zoo
AFFRICATES	
/dZ/	jeer
/tS/	cheer

continued/.....

BRITISH RP-SAMPA	WORD EXAMPLE
SHORT VOWELS	
/@/	ago
/{/	bat
/E/	bet
/I/	bit
/Q/	cod
/U/	good
/V/	bud
LONG VOWELS	
/3/	bird
/A/	bard
/i/	bead
/O/	bore
/u/	boot
DIPHTHONGS	
/@U/	zero
/aI/	pie
/aU/	cow
/E@/	hair
/eI/	pay
/I@/	peer
/OI/	boy
/U@/	contour
OTHER	
/#:/	Silence
/#/	Word boundary

As the speech synthesiser is conventional, it will not be described further here.

The phoneme list is received by a translation unit 4 which will be described in more detail below. It serves to generate, from the phoneme list, a series of command signals specifying the mouth articulation required of the face in order that it should move in a manner corresponding to the phoneme list and hence to the speech signal generated by the synthesiser 3.

These command signals are received by an image synthesis unit 5. This unit has stored in it a single video frame or bit-map image of a still picture of the desired face, and serves to generate a continuous video signal showing this face, but with movement. Obviously this video signal can be to any standard wished; here a System I signal at 25 frames per second is assumed. The movement is generated by with the aid of a three-dimensional wire frame model. A typical such model is shown in Figure 2, with the mouth area being shown enlarged in Figure 3. It has a number of points (vertices) in three-dimensional space and lines joining these vertices define triangular areas referred to as facets. In the actual apparatus, the model exists as a set of stored data, namely, for each vertex, a vertex number and its x, y, z co-ordinates and, for each facet, a facet number and the numbers of the three vertices forming the corners of the facet. During an initialisation phase, the unit 5 determines a mapping between each facet of this reference model and a corresponding area of the bit-map image. Movement is created by repeatedly defining a changed model in which one or more of the vertices assumes a different position from the position it occupied in the reference model. The unit 5 then needs to generate a new two-dimensional bit-map image. This it does by identifying any facet of the changed model one or more of the vertices of which have moved relative to the reference model; for each such facet it employs an interpolation process in which that triangular area of the original bit-map which, in accordance with the mapping, corresponds to it is moved and/or distorted to occupy in the new bit-map image a triangular area which, in accordance with this mapping, corresponds to the facet of the changed model. Such a new bit-map image is generated for each frame of the output signal (i.e. every 40 ms). For more details of the operation and implementation of the image synthesis unit 5, reference may be made to W.J. Welsh, S. Searby and J. B.

Waite, "Model Based Image Coding", Br. Telecom Technol. J., vol 8, No. 3, July 1990.

The commands needed to drive the image synthesis unit 5 could, in principle, consist of sending to the unit, every 40 ms, the number of each vertex whose position differs from the reference model, accompanied by its new co-ordinates. In the interests of speed of operation, however, the unit 5 contain a stored set of action units, each of which is a data entry consisting of:

- an action unit number (e.g. 0 to 255) (1 byte)
- the number of vertices affected by the action unit
- for each such vertex:
  the vertex number (2 bytes)
  its x-co-ordinate displacement from its position in the reference model (2 bytes)
  its y-co-ordinate displacement from its position in the reference model (2 bytes)
  its z-co-ordinate displacement from its position in the reference model (2 bytes).

(If preferred, of course, x, y, z shifts relative to the previous frame could be used).

Each command may then consist simply of an action unit number followed by a scaling factor (e.g. from 0 to 255) to vary the amount of movement specified by the action unit; or if desired may contain several (in a prototype, up to five were permitted). The unit 5, upon receipt of the command, looks up the action unit(s) and uses the stored co-ordinate shifts (scaled as appropriate) for the specified vertices. If the command contains two action units both of which specify a displacement of a particular vertex, then the displacement is simply the vector sum of the two displacements.

Returning now to examine the operation of the translation unit 4, it is convenient to introduce the concept of a viseme. Just as spoken words may be regarded as composed of elemental units called phonemes, visual speech may be regarded as composed of visemes - the minimal units of visual speech, or "the smallest perceptible unit of the visual articulatory unit". Basically, a viseme is a mouth shape; the task of the translation unit is to determine what visemes are

required, and the time instants at which they occur (quantised to multiples of 40 ms), and then to generate commands at 40 ms intervals such as to generate the required visemes at the required intervals and to generate appropriate intermediate shapes for the intervening frames.

Central to the operation of the translation unit is the notion that there is not a 1:1 correspondence between phonemes and visemes. Firstly, some phonemes are visually similar or even indistinguishable; for example the consonants /p/ and /b/ are visually identical since they differ only by the degree of voicing and the articulation of the vocal tract is the same. Thus phonemes can be grouped, with phonemes of the same group being considered identical as far as viseme generation is concerned. Various groupings are possible; a typical grouping is shown in Table 2 below:

TABLE 2

Phonemes	Group
p, b, m	Consonant group 1
f, V	Consonant group 2
D, T	Consonant group 3
s, z	Consonant group 4
S, Z	Consonant group 5
k, g, N	Consonant group 6
t, d, l, n, r	Consonant group 7
w, U, u, O	"Both" group
Q, V, A	Vowel group 1
3, i, j	Vowel group 2
@, E, I, {	Vowel group 3

(note that diphthongs are absent since these are divided into their constituent vowels before processing)

Secondly, whilst it is possible to define an association between a vowel sound and a mouth shape, this is not so with a consonant where the mouth shape varies in dependence upon nearby phonemes, especially nearby vowel phonemes. In the present embodiment mouth shapes are associated both with vowels and with combinations of a consonant and a phoneme. There are a significant number

of transitions involving consonants. However a first simplification can be made by observing that a consonant to consonant transition is heavily influenced by the following vowel (or, at the end of a word before a pause, the preceding vowel) and whilst the second consonant of the two has some effect this is quite subtle and can be ignored. The present embodiment takes advantage of this by associating a consonant-vowel or vowel-consonant combination with each consonant to consonant transition. In this way, the number of mouth shapes that need be handled by the system is kept low.

To illustrate the operation of the present embodiment by way of example, if the text-to-phonetic unit 2 were to receive a signal representing the word 'affluence', it would operate to output the phoneme list /#:/ /{/ /f/ /l/ /u/ /@/ /n/ /s/ /#:/ to the translation unit 4. On receiving that phoneme list the translation unit 4 would be operable to process the phoneme list to output a series of command signals. The output command signals are illustrated in Figures 4 to 7, each of which also illustrates the contents of the input phoneme list, i.e. the phonemes themselves and their duration in samples (in this example the sample rate is 8kHz).

Firstly, the output includes three command signals corresponding to the vowels in the word. These are shown in Figure 4 where, in the lower diagram, the vowels /{/, /u/ and /@/ have been identified and are each marked with a bar indicating that the viseme allocated to that vowel has been determined; it is taken to occur at the mid-point of the vowel.

The output further includes command signals specifying the mouth shapes associated with the vowel-consonant and consonant-vowel transitions; this is illustrated in figure 5 where bars show the mouth shapes at the vowel-consonant or consonant-vowel boundaries. This leaves the consonant-consonant transitions. As mentioned earlier, the transition is regarded as being characterised chiefly by the first consonant and the next vowel following; thus the transition /f/ to /l/ is represented (in Figure 6) as the mouth shape for the consonant-vowel combination /f/ to /u/. The /n/ to /s/ transition has no following vowel and therefore the mouth shape used is that corresponding to the /@/ to /s/ vowel-consonant combination - i.e. using the preceding vowel. The preceding and following silence periods /#:/ are of course represented by a face with a closed mouth - i.e. the reference wire frame model.

At the time instants marked with the bars in Figure 6 (or rather, at the nearest 40ms period to those instants), the translation unit 4 sends to the image synthesis unit 5 a command specifying an action unit and scaling factor appropriate to the mouth shape in question. At 40 ms intervals between those instants, it is necessary to send a command specifying a mouth shape intermediate the two mouth shapes. For example, between the instant marked {f and the instant marked fu it sends a command specifying the two action units corresponding to the vowel-consonant combination /{/ to /f/ and the consonant-vowel combination /f/ to /u/ respectively, albeit with reduced scaling factors, so as to achieve a smooth transition between the two shapes. Thus at a point x% of the way between the two instants, the action unit for the combination /{/ to /f/ would be sent with a scale factor 1-x/100 times its scale factor at the {f point, along with the action unit for the combination /f/ to /u/ with a scale factor of x/100 times its scale factor at the fu point. Figure 7 shows this process graphically. It will be seen that for the purposes of creating intermediate command signals, the mouth shape associated with the silence phoneme is not affected by the following mouth shape before the centre of the silence phoneme is reached.

Of the 11 groups in the above Table 2 above, there are 7 consonant groups, three vowel groups and one so-called "both" group. The "both" group includes both vowel phonemes and consonant phonemes. Thus, ignoring transitions involving silence, all the required vowels and vowel-consonant and consonant-vowel combinations can be represented by the vowel groups and vowel group-consonant group and consonant group-vowel group combinations shown in Table 3 below:

TABLE 3

Vowels	4
Consonant group to vowel group combinations	21
Vowel group to consonant group combinations	21
Both group to other group combinations	10
Other group to both group combinations	10
Both group to both group combinations	2
Total	68

Some of these 68 vowel groups and group combinations correspond to identical mouth shapes; moreover some mouth shapes are similar to others, differing primarily in proportions - i.e. they can be created by the same action unit but with a different scaling factor. During determination of the action units (to be described below) it was found that these 68 vowel groups and group combinations could be represented by eleven action units and an appropriate scaling factor. Table 4 below sets out these, with a description of the action unit, a note of the feature which increases with the scaling factor, and a list of the vowel groups and group combinations which can be represented by that action unit. The scaling factors to be used in creating the respective mouth shapes that correspond to given vowel groups and group combinations are also shown.

It will be realised by those skilled in the art that a larger number of action units might be defined, with the vowel groups and group combinations being more finely divided amongst the action units.

TABLE 4

Action Unit No.	Description	Vowel group or Group Combination	Scale
1	Rounded mouth, protruding lips, teeth together. Mouth shape gets more rounded with amount.	Vowel group 1 to Consonant group 5	125
		Vowel group 2 to Consonant group 5	130
		Vowel group 3 to Consonant group 5	125
		"Both" group to Consonant group 5	120
		Consonant group 5 to Vowel group 1	120
		Consonant group 5 to Vowel group 2	120
		Consonant group 5 to Vowel group 3	125
		Consonant group 5 to "Both" group	120
2	No teeth, very rounded external lip line, gap between lips straight but small. Mouth shape gets more rounded with amount.	"Both" group to Vowel group 2	150
		"Both" group to Vowel group 3	150
		"Both" group to "Both" group	150
		"Both" group to "Both" group	130
		"Both" group to Consonant group 7	120
		Consonant group 7 to "Both" group	120
3	Long mouth shape, top teeth only, bottom lip tucked. Teeth become prominent with amount.	Vowel group 1 to Consonant group 2	100
		Vowel group 2 to Consonant group 2	110
		Vowel group 3 to Consonant group 2	115
		"Both" group to Consonant group 2	100
		Consonant group 2 to Vowel group 1	100
		Consonant group 2 to Vowel group 2	100
		Consonant group 2 to Vowel group 3	115
		Consonant group 2 to "Both" group	100

4	Mouth shape long and rounded, no teeth, gap between lips round. Gap between lips gets bigger with amount.	Vowel group 1 to "Both" group Vowel group 1 to Consonant group 1 Vowel group 2 to "Both" group Vowel group 3 to Consonant group 1 "Both" group "Both" group to Vowel group 1 "Both" group to Consonant group 3 "Both" group to Consonant group 7	240 130 240 130 130 240 130 130
5	As action unit 4 but top lip is much more curved.	Vowel group 1 Consonant group 1 to Vowel group 1 Consonant group 1 to "Both" group	130 95 80
6	Long mouth shape, top and bottom teeth visible but a gap between. The gap gets bigger with amount.	Vowel group 3 to Consonant group 6 "Both" group to Consonant group 6 Consonant group 1 to Vowel group 2 Consonant group 6 to Vowel group 3	110 110 130 110
7	Rounded mouth shape, top and bottom teeth visible but a gap between. The gap gets bigger with amount.	Vowel group 1 to Consonant group 6 Vowel group 3 Consonant group 6 to Vowel group 1 Consonant group 6 to "Both" group Consonant group 7 to Vowel group 1	110 140 130 110 130
9	Long, slightly rounded mouth shape, top teeth visible. Top teeth become more prominent with amount.	Vowel group 2 Vowel group 2 to Consonant group 6 Consonant group 4 to Vowel group 3 Consonant group 6 to Vowel group 2 Consonant group 7 to Vowel group 3 Consonant group 7 to Vowel group 3	160 160 170 160 170 125

11	Long mouth shape, top teeth visible. Top teeth become more prominent with amount.	Vowel group 3 to Consonant group 4 Vowel group 3 to Consonant group 7 "Both" group to Consonant group 4 Consonant group 4 to "Both" group	130 120 105 105
12	same as action unit 11 but the top lip is not as rounded.	Vowel group 1 to Consonant group 4 Vowel group 1 to Consonant group 7 Vowel group 2 to Consonant group 4 Vowel group 2 to Consonant group 7 Consonant group 4 to Vowel group 1 Consonant group 4 to Vowel group 2	100 100 120 120 130 110 120
13	Long mouth shape with top teeth and tongue. Teeth become more prominent with amount.	Vowel group 1 to Consonant group 3 Vowel group 2 to Consonant group 3 Vowel group 3 to Consonant group 3 Consonant group 3 to Vowel group 1 Consonant group 3 to Vowel group 2 Consonant group 3 to Vowel group 3 Consonant group 3 to "Both" group	105 110 115 105 105 130 105

The translation unit 4 may be implemented by means of a suitably programmed processing unit, and thus in Figure 8 is shown as comprising a processor 10, program memory 11, and a number of stores containing look-up tables. More particularly these comprise a diphthong table 12, a phoneme group table 13 and an action unit table 14. These are shown separately for clarity but of course a single memory could in practice contain the program and look-up tables. The operation of the program stored in the memory 11 will now be described in more detail with reference to the flowcharts shown in Figures 9 to 13.

The flowchart of Figure 9 simply illustrates the operation of the apparatus as a whole, setting the context within which the algorithm represented by Figures 10 to 13 occurs. The algorithm is stored in the program memory 11 and is executable to generate an action unit file (comprising action units and scaling

factors) which forms the basis for the command signals to be sent to the image synthesis unit 5. Thus following initialisation in step 100, a text message is received 102 by the text-to-phonetic unit 2 of the speech synthesiser, which produces at 104 a phoneme file. When receipt of this file is recognised by the translation unit 4 (step 106), translation takes place (step 108) of the phoneme list into an action unit file (produced at 110). This forms the basis for the command signals which are transmitted (step 112) to the image synthesis unit 5 whilst the phoneme file is sent to the synthesiser 3. If desired during silence (step 114) or during speech (step 116) additional action units may be generated to create random (or other) head movement.

The operation of step 108 begins with the expansion of diphthongs and affricates using the program steps illustrated by the flow chart shown in Figure 10. The program reads (step 120) each element of the phoneme file in turn and determines (step 122) whether that phoneme is represented by two characters. If it is, the program causes the processor (step 124) to divide the element into its constituent characters and replaces the element with the two phonemes represented by those characters. The duration of each is set to one half of the duration of the diphthong or affricate phoneme which has been split. A variable (noofphonemes) measuring the number of phonemes in the list of phonemes output is then incremented by one (step 126). Otherwise, the element is added to the phoneme list (step 128).

It will be seen how the illustrated program steps are executable to convert diphthongs such as /aI/, /aU/, and /eI/ to phoneme pairs /{/+/I/, /{/+/U/ and /E/+/I/ respectively with the aid of the diphthong table 12. Similarly, the program is executable to divide the affricates /dZ/ and /tS/ into two phonemes.

This is followed by (Figures 11A - 11D) examination of the phoneme list created by the process illustrated in Figure 10 element by element. For each element after the initial silence phoneme, a phoneme combination or vowel and associated time interval is recorded in an intermediate output file. Thus, each entry identifies the phoneme combination or vowel along with a time interval to be created between the previous mouth shape instant and the current mouth shape instant (i.e. the time interval corresponds to the distances between the bars in Figure 6). Unless stated otherwise below, after each entry, the program returns to

a decision step 180 to determine whether the last element of the phoneme list has been reached. If it has, then examination of the phoneme list ends. If it has not, the program returns to a current element classifying step 130.

In order to examine the phoneme list, for each element it is first determined whether the element is a vowel, consonant or silence (Figure 11A - step 130).

If a vowel is found in the current element classifying step 130, the steps illustrated in Figure 11B are carried out. Firstly, it is found whether the previous phoneme in the phoneme list is a silence, consonant or vowel (step 140). If the previous phoneme is a silence phoneme, then the time interval before the vowel mouth shape is set to the sum of half of the vowel duration and half of the silence duration (step 141). The silence to vowel transition is then entered into the intermediate output file together with the calculated time interval (step 142). If the previous phoneme is a vowel phoneme, then the time interval between the vowel mouth shapes is set to the sum of half of the duration of the current vowel and half of the duration of the previous vowel (step 143). Again, the vowel itself (e.g. /@/) and associated time interval are then entered into the intermediate output file (step 144). If the previous phoneme is a consonant phoneme, then it is determined whether the phoneme before the previous phoneme is a silence (step 145). If it is, then the time interval from the previous mouth shape is set to the half the duration of the current vowel (step 146) and the vowel is entered into the intermediate output file together with the calculated time interval (step 147). If it is not, then the time interval from the previous mouth shape is set to the duration of the consonant (step 148) and the vowel to consonant combination (e.g. /I/ to /u/) and the associated time interval (step 149) are entered into the intermediate output file. At this point the program does not return to the decision step 180, but causes a further entry to be made in the transition file (steps 146,147), the entry containing a time interval equal to half the duration of the current vowel and the vowel itself (e.g. /u/).

One effect of the steps of Figure 11B is to ensure that the mouth shape corresponding to the current vowel coincides with the middle of the vowel phoneme.

If a silence is found in the current phoneme classifying step (130), then the steps of Figure 11C are carried out. Firstly, it is found whether the previous phoneme in the phoneme list is a silence, consonant or vowel (step 150). If the previous phoneme is a silence, then an error is indicated (step 151). If the silence is preceded by a vowel then a time interval from the previous mouth shape is set to the sum of half of the vowel duration and half of the silence duration (step 152), and the vowel to silence transition is recorded in the intermediate output file (step 153) together with the time interval. If the previous phoneme is a consonant then the time interval from the last mouth shape is set to sum of the duration of the consonant and half the duration of the current silence (step 154). In this case, the vowel-consonant combination to vowel transition (e.g. /@s/ to /#:/) and associated time interval is entered into the intermediate output file (step 155).

If a consonant is found in step 130, the steps illustrated in Figure 11D are carried out. Firstly, the previous phoneme is classified as a vowel, silence or consonant (step 160). If it is a vowel, then the time interval is set to half the duration of the vowel (step 161), and the vowel-consonant combination (e.g. /{/ to /f/) is recorded together with the time interval in the intermediate output file (step 162). If the previous phoneme is a consonant, then the program searches forward through the phoneme list for a vowel phoneme (step 163). If one is found the consonant-vowel combination (of the previous consonant and the later vowel) (e.g. /f/ to /u/) and the associated time interval (equal to the duration of the previous consonant) are entered in the intermediate output file (steps 164,165). If no vowel is found in the forward search (step 163) then the program causes the processor to search backwards for a vowel (step 166). If this search is successful then the vowel-consonant combination (of the earlier vowel and the current consonant - e.g. /@/ to /s/) is recorded together with an associated time interval (equal to the duration of the previous consonant) (steps 167,168). If neither a forward search nor a backward search finds a vowel an error indication results (step 169). If the phoneme immediately preceding the current consonant is found to be a silence, then a forward search for a vowel is carried out (step 170); if a vowel is found a time interval equal to the sum of the durations of the current consonant and half the duration of the preceding silence is recorded together with a silence to consonant-vowel combination transition in the intermediate output file

(steps 171,172). If no vowel is found in the word then an error is indicated (step 173).

In Figure 12, the vowels and phoneme combinations in the intermediate output file are converted into vowel groups and phoneme group combinations by accessing the look-up table 13. In principle the contents of this could be as set out in Table 2 above, so that each vowel or phoneme combination translates to a group number. However it was found more convenient to represent each group not by a group number but by one designated phoneme of the group; for example the phonemes /p/, /b/ and /m/ were all translated into /p/. To achieve this the processor is controlled by the program illustrated in Figure 12. For each element in the intermediate output file, the type of element is determined (step 190) to be one of: a vowel (steps 192 are carried out); a vowel/consonant combination (steps 194 are carried out); a vowel/silence transition (steps 196 are carried out); or a combination to silence transition (steps 198 are carried out). The steps (192,194,196,198) are effective to convert each of the constituent vowel or consonants to a vowel or consonant chosen to represent the group. This procedure returns a group/group combination list which now contains a maximum of 68 different vowel groups and phoneme group combinations as discussed above.

In Figures 13A and 13B the resulting group list is converted to an action unit file using the action unit look-up table 14 (the contents of which are as set out in columns 3, 1 and 4 of Table 3 above - or with representative phonemes in column 3 if this is the preferred option) to find the action unit representing each element in the group/group combination list. The action unit file may then be used to provide a sequence of command signals generated at 40 ms intervals.

In more detail, the conversion procedure begins with fetching the first element from the group list (step 200), whereafter the action unit look-up table is accessed to determine the action unit and scaling factor associated with that element (step 201). Then, the number of entire 40ms periods within the time interval associated with the first element is calculated (step 202). The scaling factor of the initial action unit is then divided by the number of periods to give an increment value (step 203). The procedure then enters a loop of instructions (step 204) producing a command signal for each 40ms period. The scaling factor in the

command signal is increased (from zero) by the calculated increment each time the loop of instructions is executed.

The next element in the group list is then fetched (Figure 13B - step 205), and the corresponding action unit and scaling factor is found using the action unit look-up table 14 (step 206). As in step 202, the number of whole 40ms periods within the time interval associated with that element of the group list is then found (step 207). As before, the scaling factor of the action unit associated with the current element is then divided by the number of periods calculated to give an increment value (step 208). The scaling factor of the previous element in the group list is divided by the same number to give a decrement value (step 209). The procedure then enters a loop of instructions to calculate the command signals to be output. These comprise a weighted combination of the action unit produced in the relation to the previous element, and the action unit associated with the current element in the group list. The weight given to the previous action unit is decreased by decrementing the scaling factor by the decrement value for each 40ms period, whereas the weight given to the current action unit is increased by increasing the scaling factor (from zero) by the increment value for each 40ms period (step 210). In this way the command signals output provide a stepped transition from one mouth shape to the next.

Similar operations (steps 206 to 210) are then applied to each subsequent element in the group list until a termination element is reached.

Command signals are generated on the basis of the action unit file and are transmitted to the image synthesis unit 5 at 40ms intervals to enable the generation of an image of a head which has articulation corresponding to the output of the text-to-speech synthesiser.

It will be noted from the above discussion that the viseme or mouth shape chosen for a vowel is one allocated in advance to that vowel, that the mouth shape chosen for a vowel-consonant (or vice versa) combination is one allocated in advance to that combination, and the mouth shape chosen for a consonant-consonant transition is one allocated in advance to the first of the consonants in the same context - i.e. in this example, having the same following (or, in default, preceding) vowel. If desired - albeit with the penalty of increasing the number of action units required - the selection of mouth shapes may be made more context

dependent. For example, one might choose a mouth shape for a consonant-vowel transition with the choice being dependent not only on the consonant and on the following vowel but also on the preceding vowel (i.e. consonant-vowel-consonant combination). The choice for a consonant-consonant transition could be made to depend on the first consonant and both the following and preceding vowels (if present) or indeed on the two consonants and two vowels.

Little has been said so far about how the action units stored in the image synthesis unit 5 are generated. This in the prototype was accomplished by making a video recording of a person speaking words containing all of the required 68 vowel groups and vowel group/consonant group combinations, and using a frame grabber to display still frames of the recording so that those frames corresponding to the vowels and those frames corresponding to consonant/vowel combinations could be identified manually. Once those frames (in bitmap form) had been identified it was then necessary to determine the displacements of from the reference wire frame model which those frames represented. This is done using a conformation program which calculates the deformation required of a wire frame model to fit a given bitmap image.

CLAIMS

1. A method of generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, comprising:

receiving a sequence of phonetic representations corresponding to successive portions of the utterance;

identifying a mouth shape for each phonetic representation of a first type;

identifying a mouth shape for each transition from a phonetic representation of the first type to a phonetic representation of a second type, for each transition from a phonetic representation of the second type to a phonetic representation of a first type and for each transition from a phonetic representation of the second type to a phonetic representation of the second type; and

generating a sequence of image frames including the identified shapes.

2. A method of generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, comprising:

receiving a sequence of phonetic representations corresponding to successive phonemes of the utterance;

identifying a mouth shape for each vowel phoneme;

identifying a mouth shape for each transition from a vowel phoneme to a consonant phoneme, for each transition from a consonant phoneme to a vowel phoneme and for each transition from a consonant phoneme to a consonant phoneme; and

generating a sequence of image frames including the identified shapes.

3. A method according to claim 2 in which the identification of a mouth shape for each transition between consonant and vowel phonemes is performed as a function of the vowel phoneme and the consonant phoneme.

4. A method according to claim 2 or 3 in which the identification of a mouth shape for each transition between two consonant phonemes is performed as a function of the first of the two consonant phonemes and of the vowel phoneme which most closely follows or precedes it.

5. A method according to claim 2 or 3 in which the identification of a mouth shape for each transition between two consonant phonemes is performed as a function of the first of the two consonant phoneme and of the vowel phoneme which most closely follows it or in the absence thereof that which precedes it.

6. A method according to claim 3, 4 or 5 in which the identification is performed as a function of only those phonemes specified in that claim.

7. A method according to claim 3, 4 or 5 in which the identification is performed as a function also of at least one other phoneme within the same word.

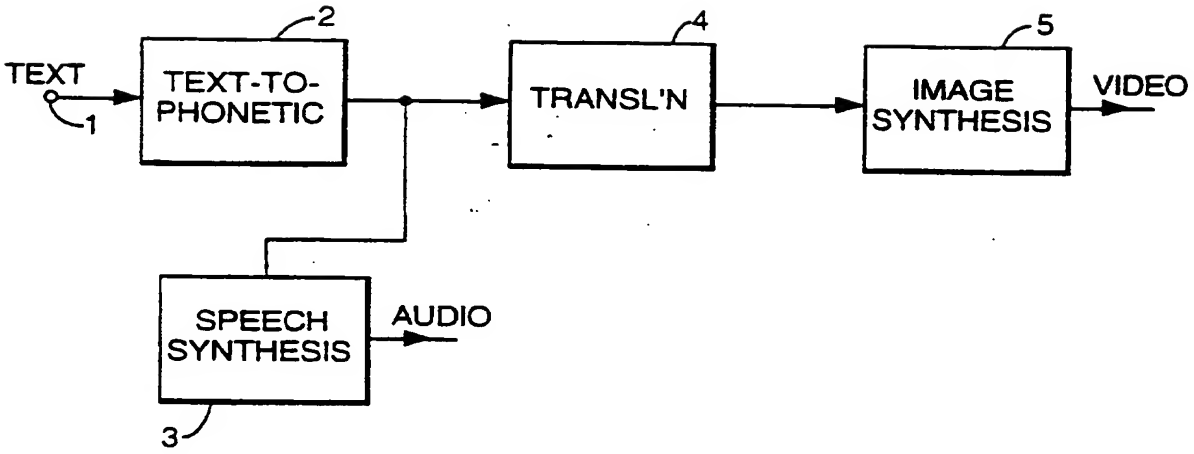
8. A method according to any one of the preceding claims including generating for each identified mouth shape a command specifying that shape and generating intermediate commands each of which specifies a shape intermediate the shapes specified by the preceding and following commands.

9. An apparatus for generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, comprising:
means arranged in operation to receive a sequence of phonetic representations corresponding to successive portions of the utterance and in response thereto to
identify a mouth shape for each phonetic representation of a first type and
identify a mouth shape for each transition from a phonetic representation of the first type to a phonetic representation of a second type, for each transition from a phonetic representation of the second type to a phonetic representation of a first type and for each transition from a phonetic representation of the second type to a phonetic representation of the second type;
and means for generating a sequence of image frames including the identified shapes.

10. A method of generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, substantially as herein described with reference to the accompanying drawings.

11. An apparatus for generating signals representing a moving picture of a face having visible articulation matching a spoken utterance, substantially as herein described with reference to the accompanying drawings.

Fig.1.


TEXT
1
2
TEXT-TO-PHONETIC
4
TRANSL'N
5
IMAGE SYNTHESIS
VIDEO
SPEECH SYNTHESIS
AUDIO
3

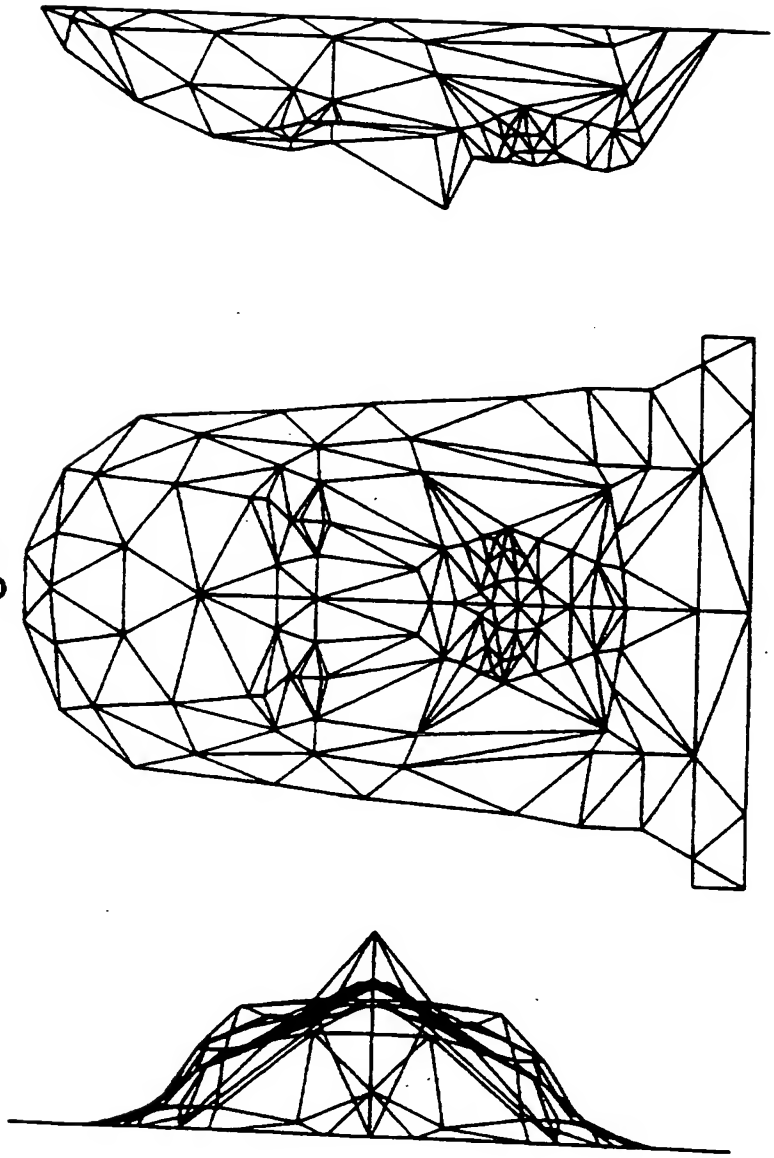


Fig.2.

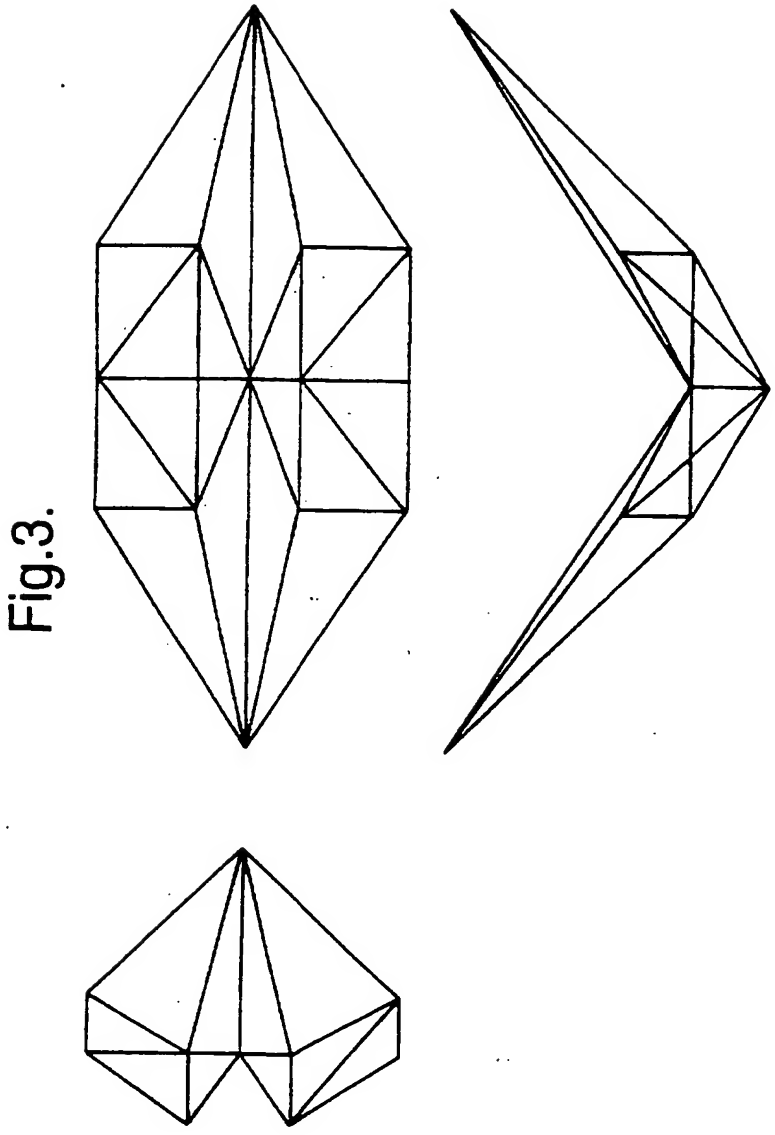


Fig.3.

Fig.4.

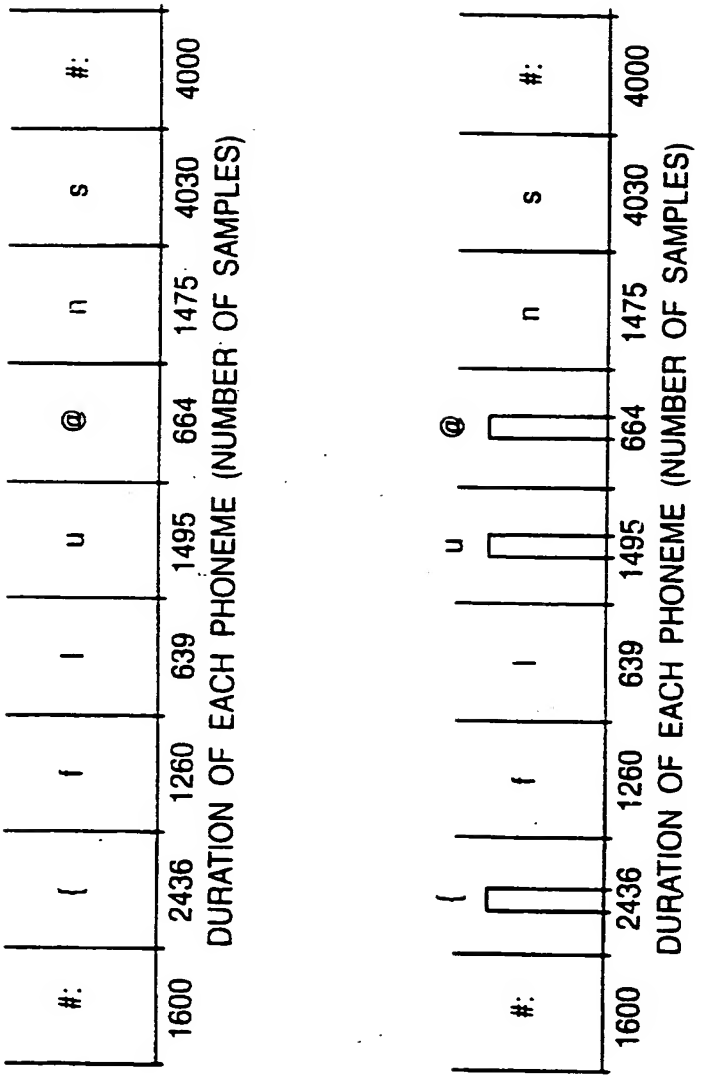

#:
{
f
I
u
@
n
s
#:
1600
2436
1260
639
1495
664
1475
4030
4000
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)
#:
{
f
I
u
@
n
s
#:
1600
2436
1260
639
1495
664
1475
4030
4000
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)

Fig.5.

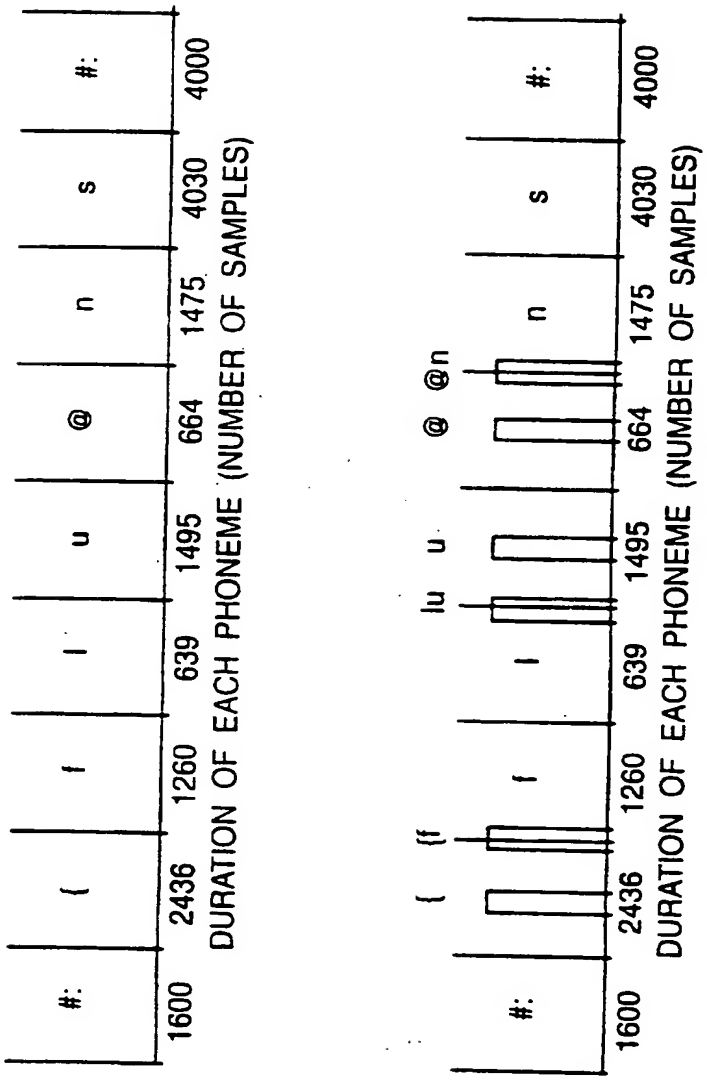

#:
{
f
I
u
@
n
s
#:
1600
2436
1260
639
1495
664
1475
4030
4000
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)
{ ff
lu u @ @n
#:
1600
2436
1260
639
1495
664
1475
4030
4000
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)

Fig.6.

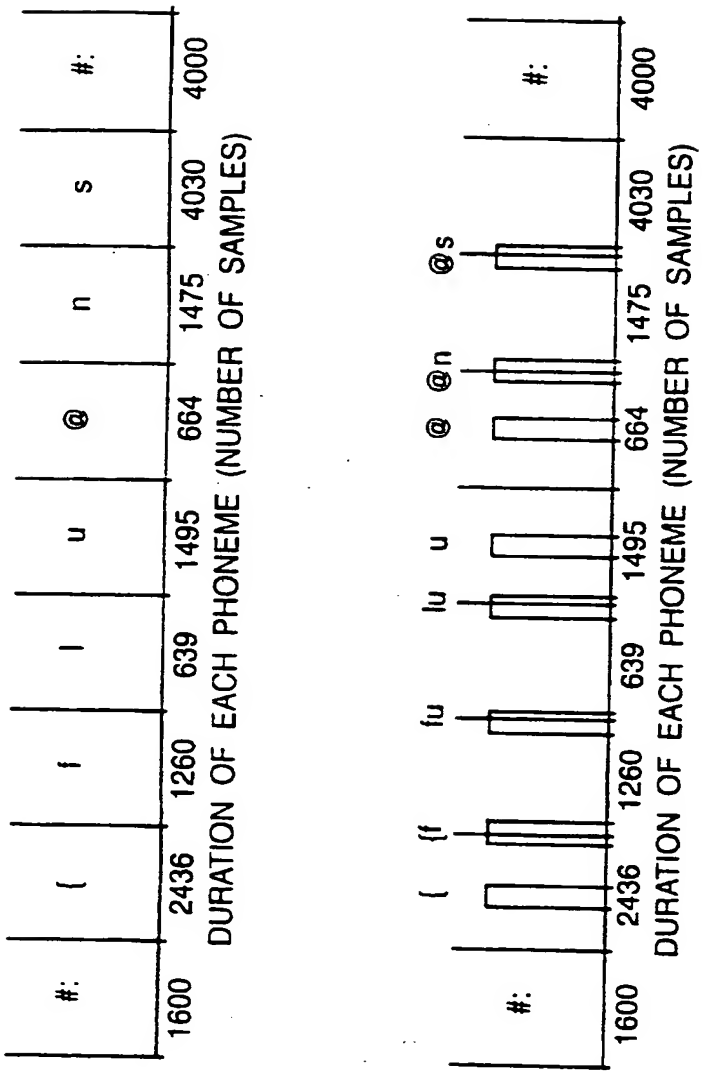
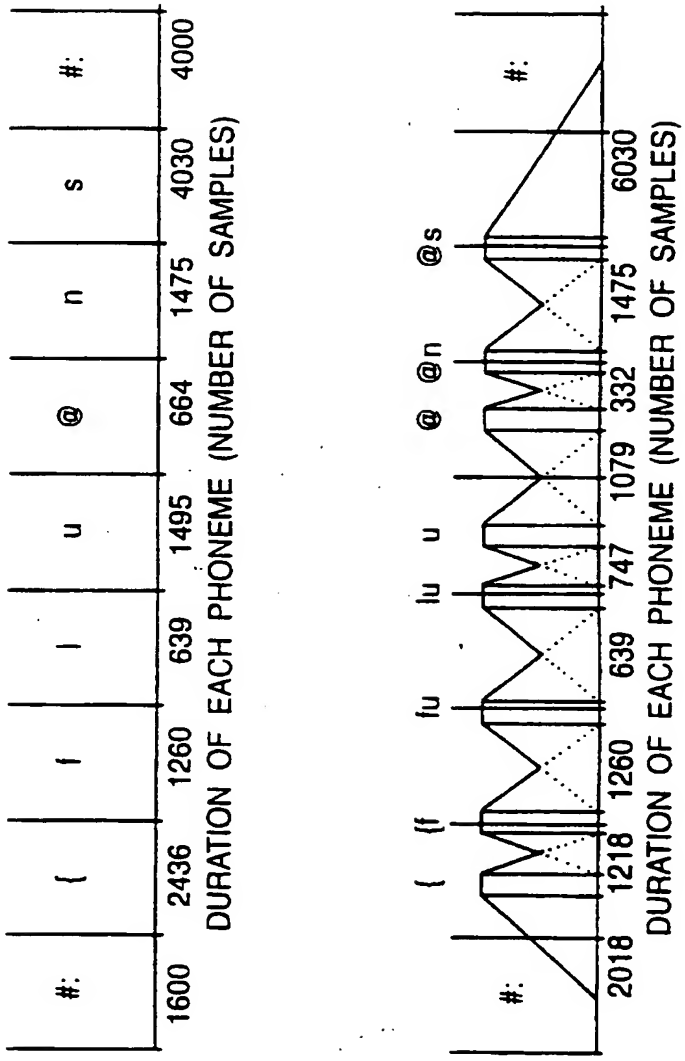

#:
(
f
I
u
@
n
s
#:
1600
2436
1260
639
1495
664
1475
4030
4000
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)
#:
(
ff
fu
lu
u
@
@n
@s
#:
1600
2436
1260
639
1495
664
1475
4030
4000
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)

Fig.7.


#: { f I u @ n s #:
1600 2436 1260 639 1495 664 1475 4030 4000
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)
#: { f fu lu u @ @n @s #:
2018 1218 1260 639 747 1079 332 1475 6030
DURATION OF EACH PHONEME (NUMBER OF SAMPLES)

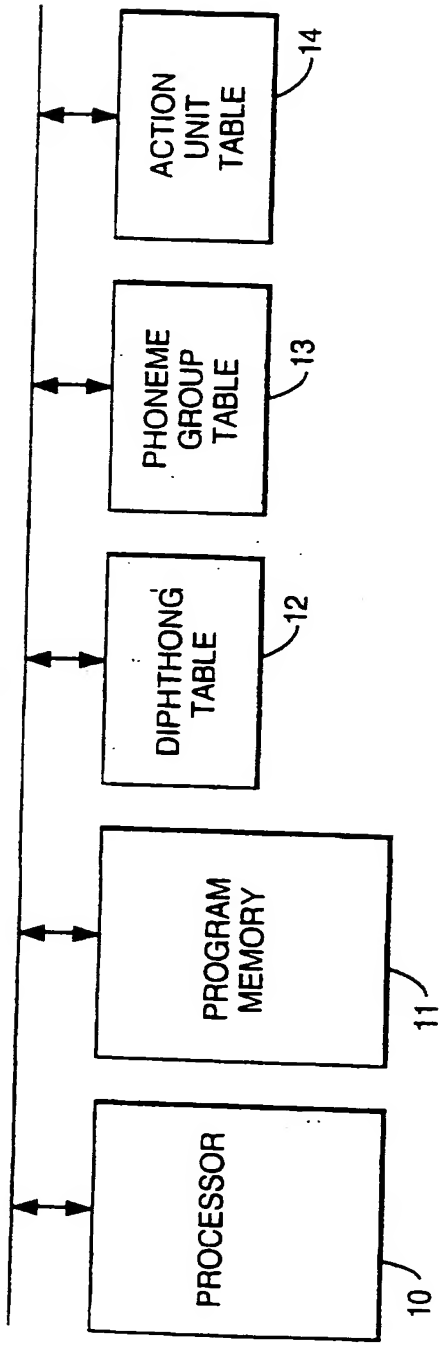

ACTION UNIT TABLE
14
PHONEME GROUP TABLE
13
DIPHTHONG TABLE
12
PROGRAM MEMORY
11
PROCESSOR
10

Fig.8.

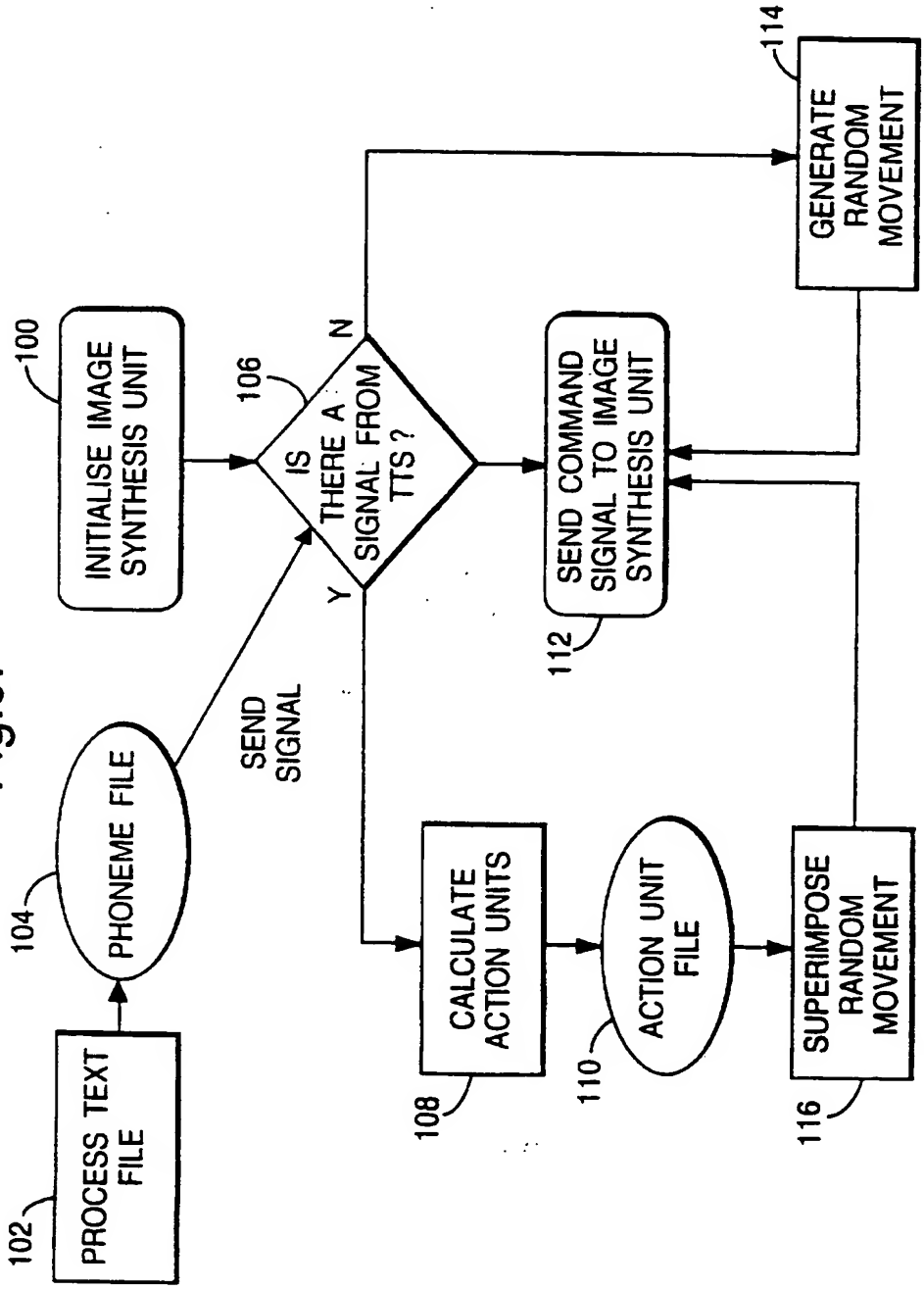

INITIALISE IMAGE SYNTHESIS UNIT
100
IS THERE A SIGNAL FROM TTS ?
106
N
Y
GENERATE RANDOM MOVEMENT
114
SEND COMMAND SIGNAL TO IMAGE SYNTHESIS UNIT
112
SEND SIGNAL
PHONEME FILE
104
PROCESS TEXT FILE
102
CALCULATE ACTION UNITS
108
ACTION UNIT FILE
110
SUPERIMPOSE RANDOM MOVEMENT
116

Fig.9.

Fig.10.

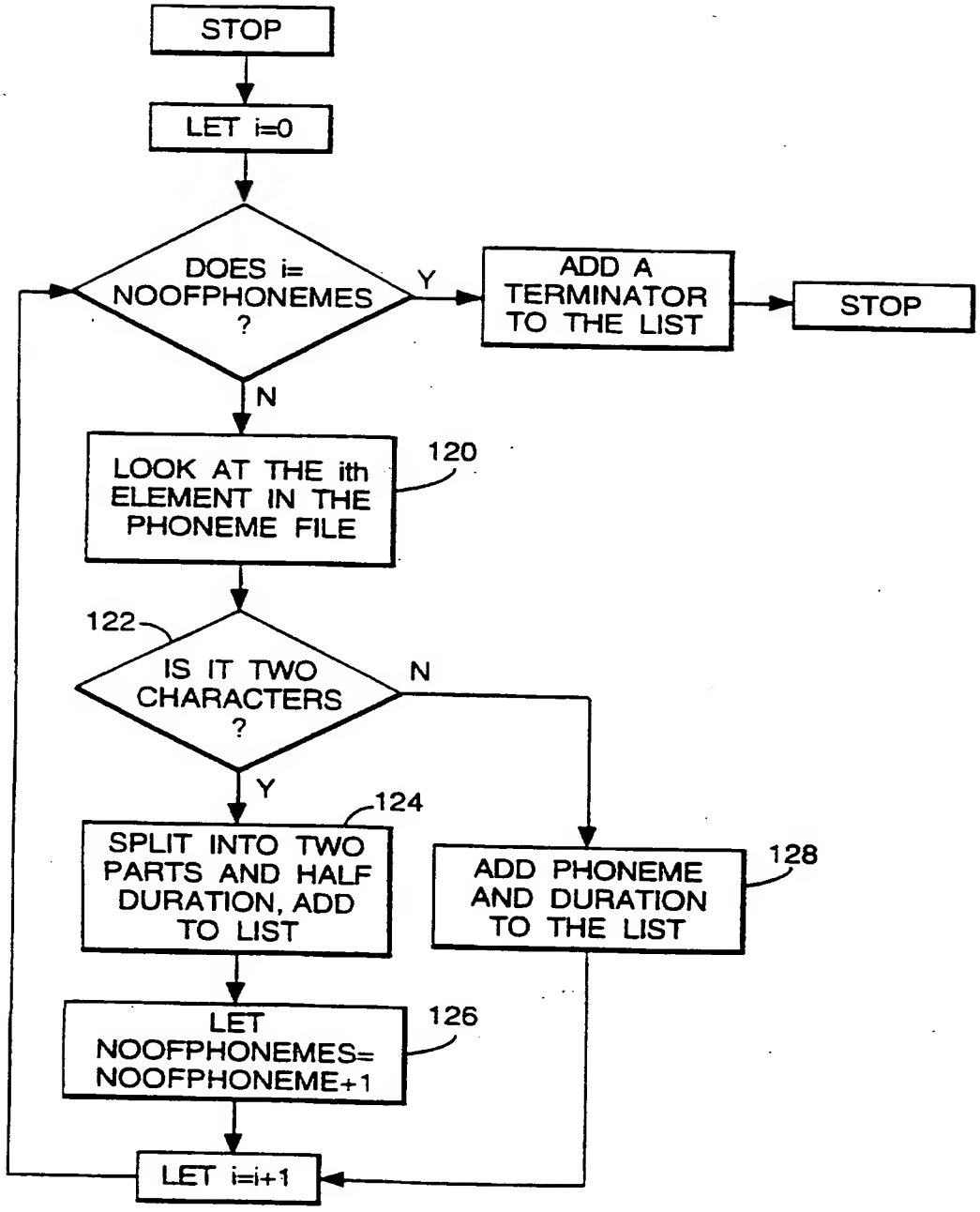
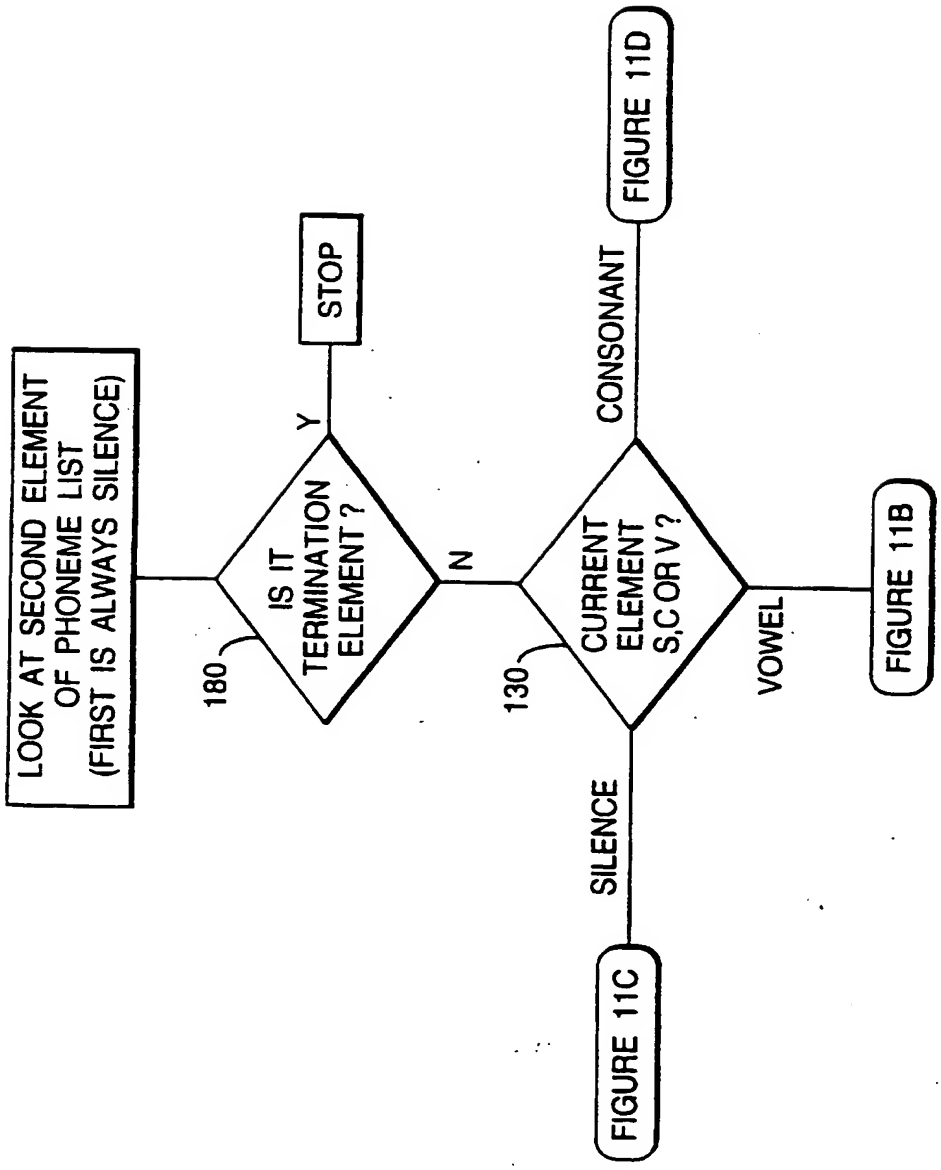

STOP
LET i=0
DOES i= NOOFPHONEMES ?
Y
ADD A TERMINATOR TO THE LIST
STOP
N
LOOK AT THE ith ELEMENT IN THE PHONEME FILE
120
122
IS IT TWO CHARACTERS ?
N
Y
124
SPLIT INTO TWO PARTS AND HALF DURATION, ADD TO LIST
128
ADD PHONEME AND DURATION TO THE LIST
LET NOOFPHONEMES= NOOFPHONEME+1
126
LET i=i+1

Fig.11A.


LOOK AT SECOND ELEMENT OF PHONEME LIST (FIRST IS ALWAYS SILENCE)
180
IS IT TERMINATION ELEMENT ?
Y
STOP
N
130
CURRENT ELEMENT S,C OR V ?
CONSONANT
FIGURE 11D
VOWEL
FIGURE 11B
SILENCE
FIGURE 11C

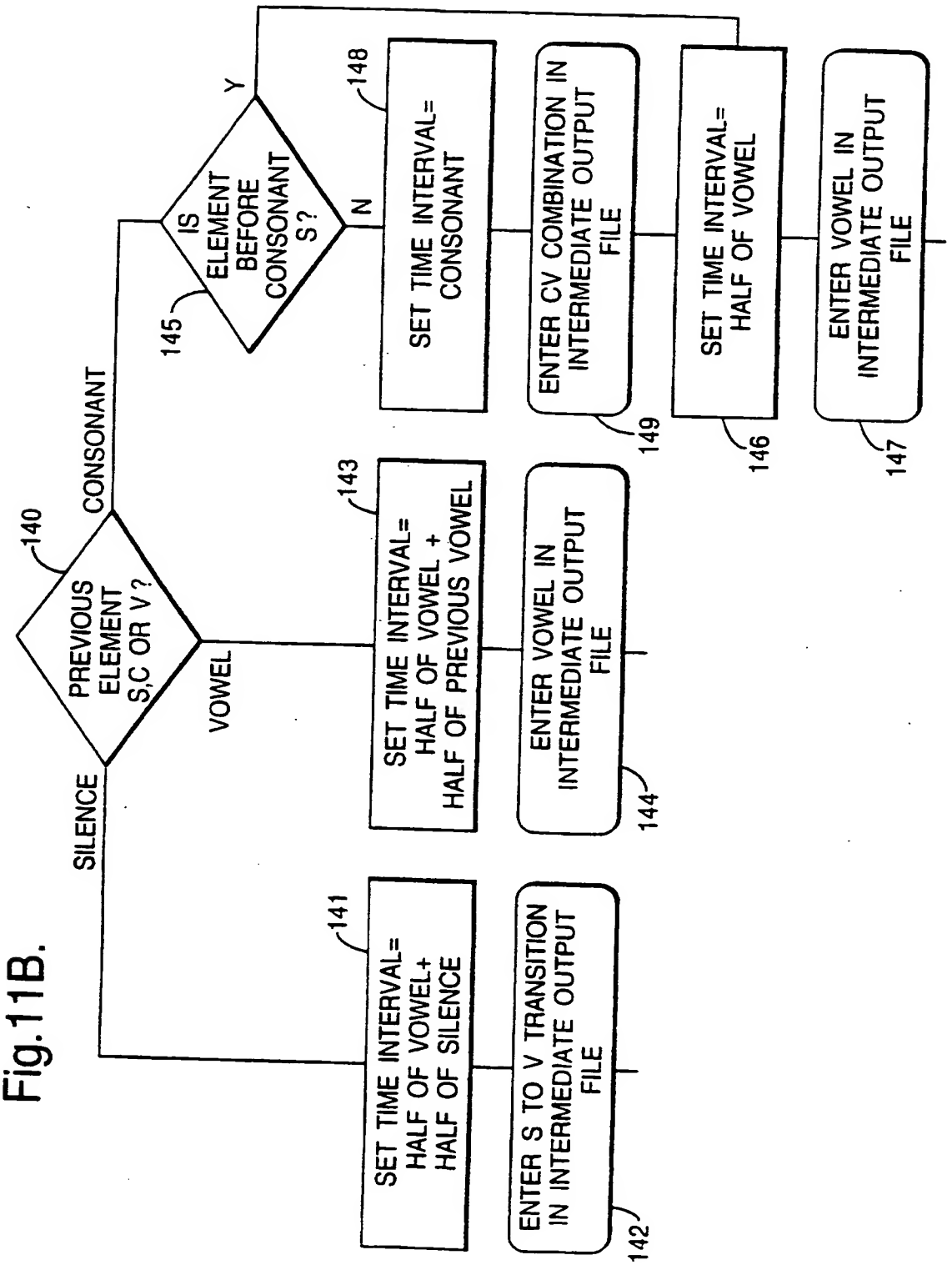

PREVIOUS ELEMENT S,C OR V ?
140
CONSONANT
VOWEL
SILENCE
IS ELEMENT BEFORE CONSONANT S?
145
Y
N
SET TIME INTERVAL= HALF OF VOWEL + HALF OF SILENCE
141
ENTER S TO V TRANSITION IN INTERMEDIATE OUTPUT FILE
142
SET TIME INTERVAL= HALF OF VOWEL + HALF OF PREVIOUS VOWEL
143
ENTER VOWEL IN INTERMEDIATE OUTPUT FILE
144
SET TIME INTERVAL= CONSONANT
148
ENTER CV COMBINATION IN INTERMEDIATE OUTPUT FILE
149
SET TIME INTERVAL= HALF OF VOWEL
146
ENTER VOWEL IN INTERMEDIATE OUTPUT FILE
147

Fig.11B.

Fig.11C.

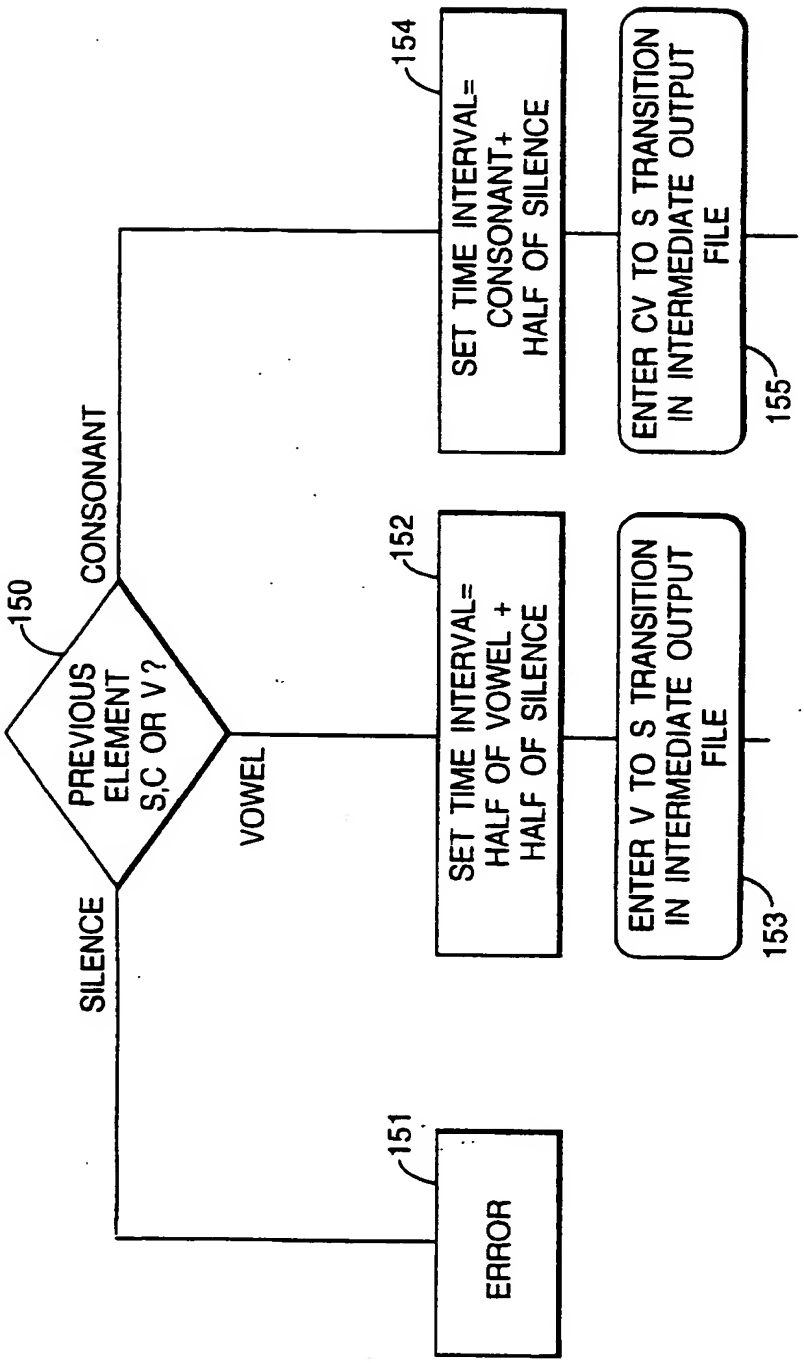

150
PREVIOUS ELEMENT S,C OR V ?
CONSONANT
VOWEL
SILENCE
154
SET TIME INTERVAL= CONSONANT+ HALF OF SILENCE
155
ENTER CV TO S TRANSITION IN INTERMEDIATE OUTPUT FILE
152
SET TIME INTERVAL= HALF OF VOWEL + HALF OF SILENCE
153
ENTER V TO S TRANSITION IN INTERMEDIATE OUTPUT FILE
151
ERROR

Fig.11D.

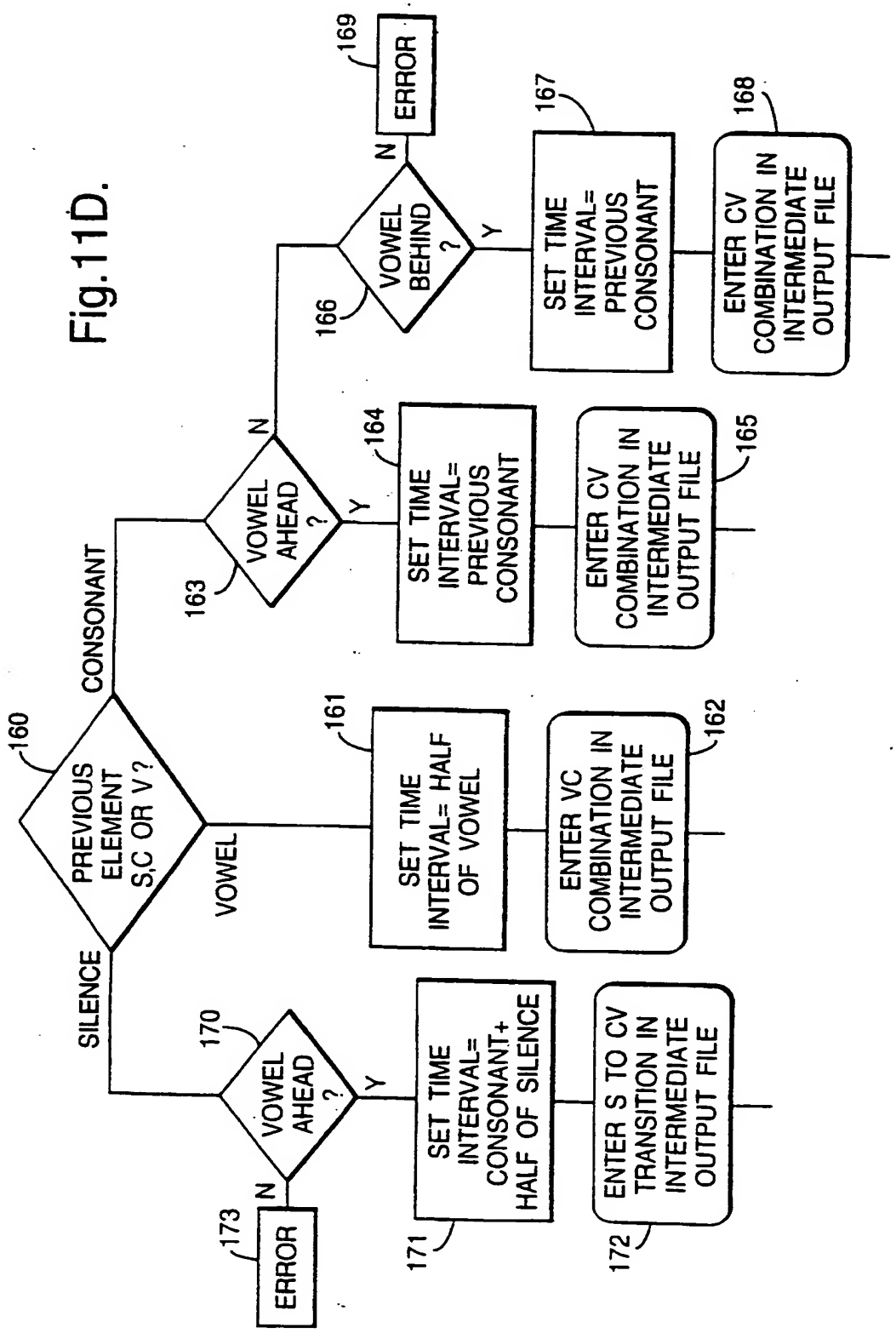

PREVIOUS ELEMENT S,C OR V ?
160
CONSONANT
VOWEL
SILENCE
VOWEL AHEAD ?
163
N
Y
SET TIME INTERVAL= PREVIOUS CONSONANT
164
ENTER CV COMBINATION IN INTERMEDIATE OUTPUT FILE
165
VOWEL BEHIND ?
166
N
Y
ERROR
169
SET TIME INTERVAL= PREVIOUS CONSONANT
167
ENTER CV COMBINATION IN INTERMEDIATE OUTPUT FILE
168
SET TIME INTERVAL= HALF OF VOWEL
161
ENTER VC COMBINATION IN INTERMEDIATE OUTPUT FILE
162
VOWEL AHEAD ?
170
N
Y
ERROR
173
SET TIME INTERVAL= CONSONANT+ HALF OF SILENCE
171
ENTER S TO CV TRANSITION IN INTERMEDIATE OUTPUT FILE
172

Fig.12.

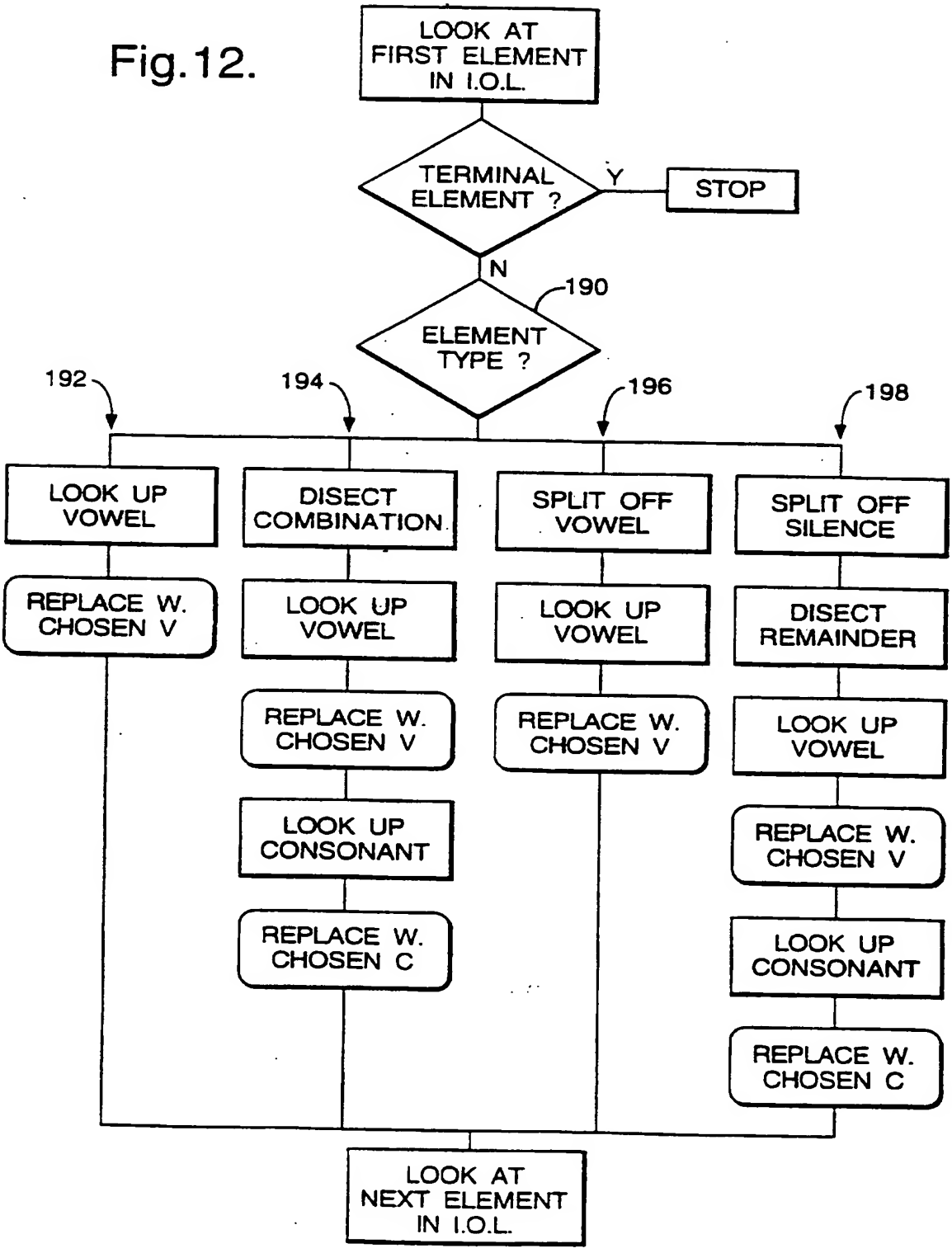

LOOK AT FIRST ELEMENT IN I.O.L.
TERMINAL ELEMENT ?
Y
STOP
N
190
ELEMENT TYPE ?
192
194
196
198
LOOK UP VOWEL
REPLACE W. CHOSEN V
DISECT COMBINATION
LOOK UP VOWEL
REPLACE W. CHOSEN V
LOOK UP CONSONANT
REPLACE W. CHOSEN C
SPLIT OFF VOWEL
LOOK UP VOWEL
REPLACE W. CHOSEN V
SPLIT OFF SILENCE
DISECT REMAINDER
LOOK UP VOWEL
REPLACE W. CHOSEN V
LOOK UP CONSONANT
REPLACE W. CHOSEN C
LOOK AT NEXT ELEMENT IN I.O.L.

Fig.13A.

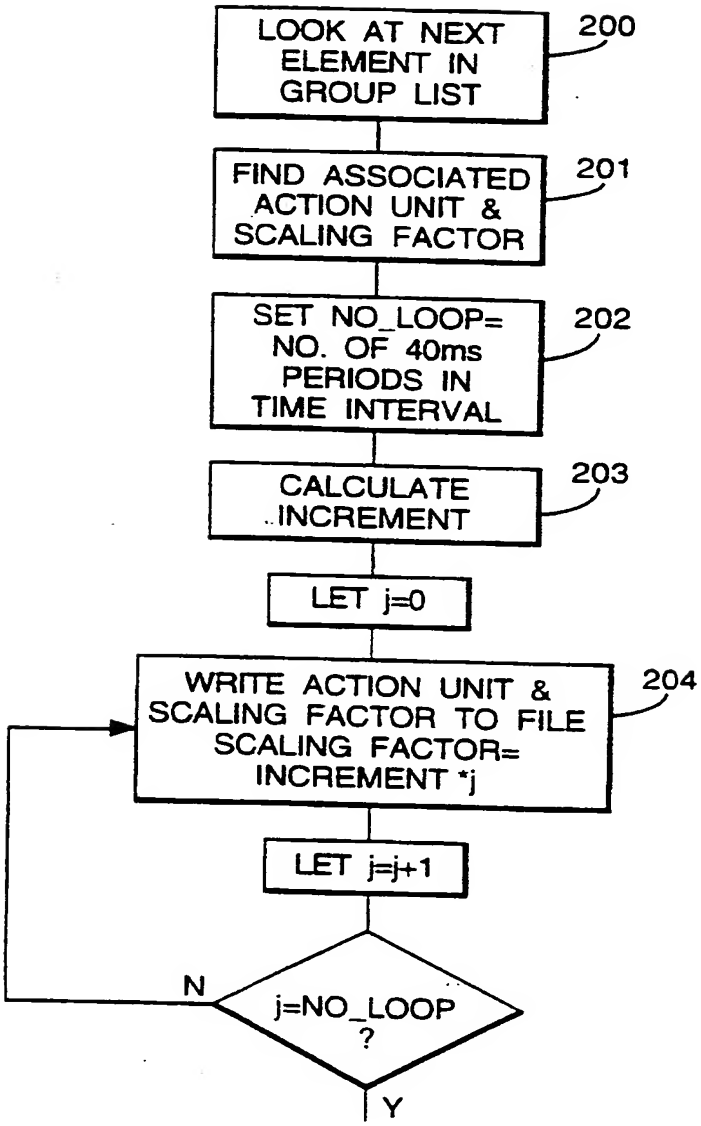
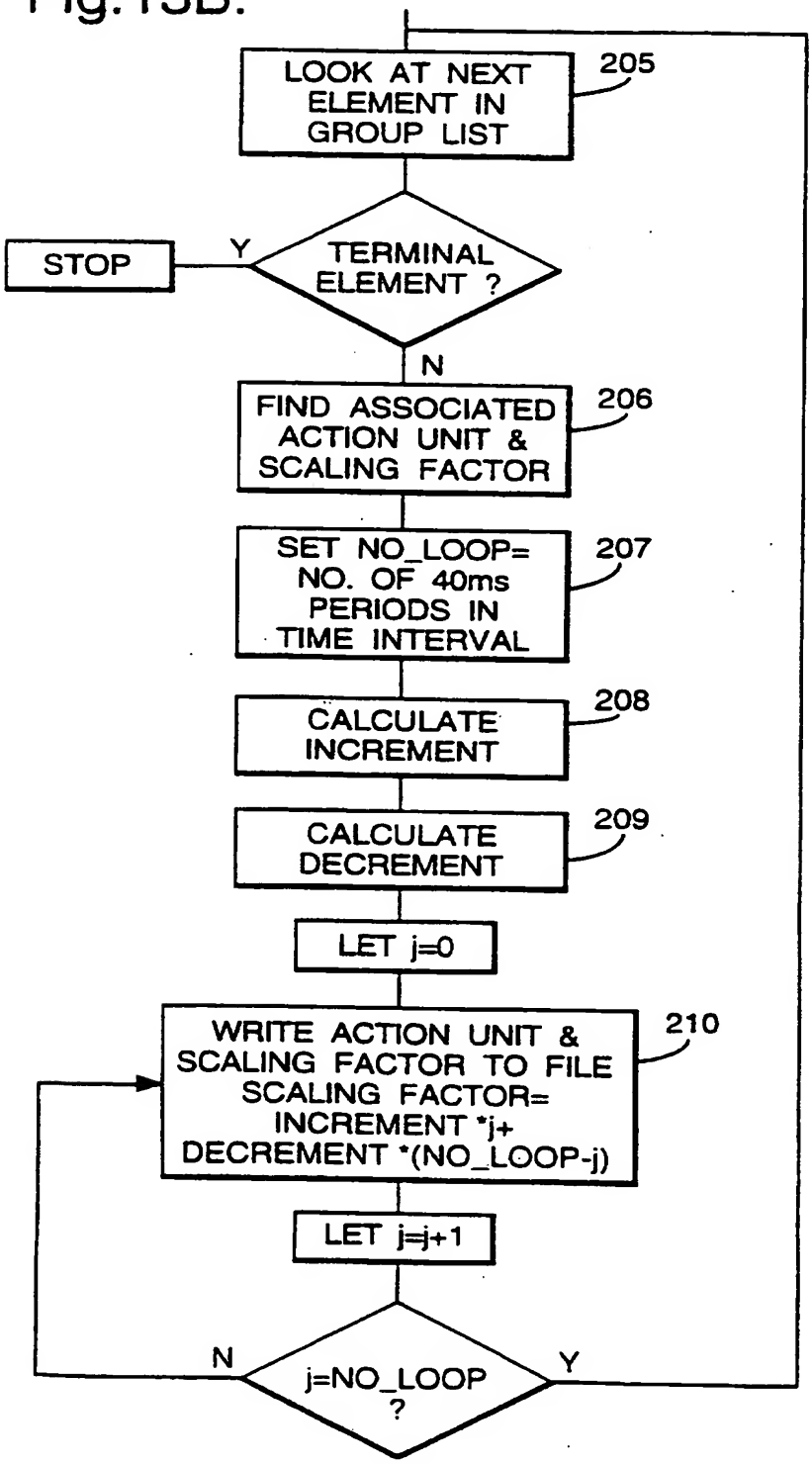

LOOK AT NEXT ELEMENT IN GROUP LIST
200
FIND ASSOCIATED ACTION UNIT & SCALING FACTOR
201
SET NO_LOOP= NO. OF 40ms PERIODS IN TIME INTERVAL
202
CALCULATE INCREMENT
203
LET j=0
WRITE ACTION UNIT & SCALING FACTOR TO FILE SCALING FACTOR= INCREMENT *j
204
LET j=j+1
j=NO_LOOP ?
N
Y

Fig.13B.


LOOK AT NEXT ELEMENT IN GROUP LIST
205
TERMINAL ELEMENT ?
Y
STOP
N
FIND ASSOCIATED ACTION UNIT & SCALING FACTOR
206
SET NO_LOOP= NO. OF 40ms PERIODS IN TIME INTERVAL
207
CALCULATE INCREMENT
208
CALCULATE DECREMENT
209
LET j=0
WRITE ACTION UNIT & SCALING FACTOR TO FILE SCALING FACTOR= INCREMENT *j+ DECREMENT *(NO_LOOP-j)
210
LET j=j+1
j=NO_LOOP ?
N
Y

INTERNATIONAL SEARCH REPORT

Intern al Application No
PCT/GB 97/00818

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G10L9/20 G06T15/70 - H04N7/26

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G10L G06T H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SYSTEMS & COMPUTERS IN JAPAN, vol. 22, no. 5, 1 January 1991, pages 50-59, XP000240754 SHIGEO MORISHIMA ET AL: "A FACIAL MOTION SYNTHESIS FOR INTELLIGENT MAN-MACHINE INTERFACE" see page 50, left-hand column, paragraph 3 - page 51, right-hand column, paragraph 1 see page 52, right-hand column, paragraph 2 - page 54, right-hand column, paragraph 3; figures 1,6; tables 1-3 ---	1-3,8-11
A	US 5 313 522 A (SLAGER ROBERT P) 17 May 1994 see column 1, line 12 - column 3, line 14; claims 1,6; figures 1,2 ---	1,2,9-11
	-/--	

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
23 June 1997	02.07.97
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+ 31-70) 340-2040, Tx. 31 651 epo nl, Fax (+ 31-70) 340-3016	Authorized officer Greiser, N

Form PCT/ISA 210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter ial Application No
PCT/GB 97/00818

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 689 362 A (AT & T CORP) 27 December 1995 see column 1, line 14 - line 17; claims 1-4; figures 1,2 see column 1, line 35 - line 44 see column 3, line 20 - column 5, line 29	1,2,9
A	GB 2 231 246 A (KOKUSAI DENSHIN DENWA CO LTD) 7 November 1990 see claims 1-5	1,2,9
A	US 4 913 539 A (LEWIS JOHN P) 3 April 1990 see column 1, paragraph 1; figure 1 see column 1, line 66 - column 2, line 41; claims 9-19	1,2,9

Form PCT/ISA 210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/GB 97/00818

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5313522 A	17-05-94	NONE	
EP 0689362 A	27-12-95	US 5608839 A	04-03-97
		CA 2149068 A	22-12-95
		JP 8023530 A	23-01-96
GB 2231246 A	07-11-90	JP 2234285 A	17-09-90
		JP 2518683 B	24-07-96
US 4913539 A	03-04-90	NONE	